The Baltic International Yearbook of
Cognition, Logic and Communication

December 2014 Volume 9: *Perception and Concepts*
pages 1-26 DOI: 10.4148/biyclc.v9i0.1082

DÁVID BITTER
Central European University

# IS LOW-LEVEL VISUAL EXPERIENCE COGNITIVELY PENETRABLE?

*A Critical Analysis of Some of the Purported Best Evidence*

ABSTRACT: Philosophers and psychologists alike have argued recently that relatively abstract beliefs or cognitive categories like those regarding race can influence the perceptual experience of relatively low-level visual features like color or lightness. Some of the proposed best empirical evidence for this claim comes from a series of experiments in which White faces were consistently judged as lighter than equiluminant Black faces, even for racially ambiguous faces that were labeled 'White' as opposed to 'Black' (Levin and Banaji 2006). The latter result is considered especially indicative of cognitive penetration, based on the reasoning that the relevant distortions were a function of lexical labeling, and hence the effect must have been mediated by categorization at the cognitive level. I argue that this reasoning is flawed, and that the assumptions on which it relies are questionable on both empirical and theoretical grounds. I propose an alternative, low-level explanation of the phenomena, which I argue is empirically more plausible and abductively preferable to the cognitive-penetration account. The upshot is that cognitively impenetrable perceptual systems may be psychologically more plastic and hence philosophically more significant than is nowadays commonly assumed.

Philosophers and psychologists alike have argued recently that the perceptual experience of relatively low-level visual features like color or lightness is susceptible to the influence of relatively abstract beliefs or concepts such as those regarding race (Collins & Olson 2014; Hugenberg & Sacco 2008; Levin & Banaji 2006; Macpherson 2012; Stokes 2013; Vetter & Newen 2014). This is a bold claim, for if visual experience is thus cognitively penetrable, then quite possibly any or at least most perceptual experience is cognitively penetrable. In turn, this has potentially important implications for issues in philosophy such as the justificatory role of perception (epistemology), the theory-dependence of observation (philosophy of science), or the continuity of perception with cognition (philosophy of mind). So whether or to what extent the above claim is warranted has great importance and relevance.

The central thesis of this paper is that there is in fact no good reason to assume that low-level visual experience is cognitively penetrable. I do not claim outright that such effects are empirically impossible. But I argue that some of the alleged best evidence indeed falls short of providing sufficient support for an argument for cognitive penetration. It bears emphasis that the evidence to be discussed is widely considered to constitute the most convincing case yet for the influence of high-level cognition on low-level perception. So, if my argumentation is correct, the burden is on defenders of cognitive penetration to argue why anyone should still assume that the posited kind of penetration actually occurs.

The outline of the paper is as follows. I first introduce the notion of cognitive penetration that is of current interest (Section 1), after which I summarize a series of psychological experiments that many see as a best-case example of cognitive penetration into color or lightness experience (Section 2). I then examine a relatively recent and prima facie powerful argument that draws on the results of these experiments. I argue that none of the premises of this argument are warranted, and hence there is plenty of room to resist the conclusion that low-level visual experience is cognitively penetrable (Section 3). I subsequently propose an alternative positive account of the evidence in terms of purely perceptual mechanisms (Section 4). The upshot is that cognitively impenetrable perceptual systems may be psychologically more plastic and hence philosophically more significant than is nowadays commonly assumed. I thus conclude by mentioning some important implications of the proposed account (Section 5).

## 1. COGNITIVE PENETRATION

Is our perception of the world influenced by how we think about the world? On the face of it, the answer would seem to be yes. For example, there is the story of the desert nomad whose desire for water leads to his hallucinating a source of water where there is none. Or what about the amputee whose refusal to accept the loss of a limb results in her experiencing pain in a body part that doesn't exist? The list of candidate examples is long. Accordingly, it might seem evident that our perceptual experience is cognitively penetrated through and through.

On closer scrutiny, though, the situation is not so simple. For example, mirages are real optical phenomena caused by atmospheric conditions. People tend to see water in a desert where there is none because light from the sky is refracted by hot air above the desert surface, thereby producing an apparent image as of a sheet of water on the ground. So the source of this image is not cognitive. Hence, mirages are not good examples of cognitive penetration.

Neither is the source of phantom limb pain cognitive. True, illusory visual feedback of a phantom hand caused by the mirror reflection of an amputee's normal hand can lead to an alleviation of phantom pain (Ramachandran et al. 1995). But this effect is presumably a function of somatosensory-motor coupling between the normal hand and the phantom hand, which in turn is plausibly explained by cognitively unmediated Hebbian (associative) learning. So phantom limb pain is not a convincing case of cognitive penetration, either.

Of course, there *are* perceptual effects the source of which is cognitive. But in many cases, the relation between the source cognitive state and the influenced perceptual state does not seem to be of the right kind. For example, perceptual effects mediated by changes in the stimuli (e.g., turning off the lights), the state of the sense organs (e.g., closing one's eyes), or the allocation of (spatial or object-based) attention are widely disregarded as genuine cases of cognitive penetration. In these cases, a cognitive source seems neither necessary, nor sufficient for the relevant effects to occur. On the other hand, the changes in the stimuli, the state of the sense organ, or the allocation of attention seem both necessary and sufficient. Accordingly, once such factors are fixed, many candidate phenomena (e.g., the perceptual switching of ambiguous pictures like the Necker cube) turn out not to be cases of cognitive penetration (Pylyshyn 1999; Raftopoulos 2001).[1]





A further issue is that many perceptual effects aren't arbitrarily sensitive to the contents of cognition. For example, selective attention might alter the impedance of the ear or the aperture of the eye. Yet such influences typically only affect the amplitude of a transducer's response, rather than the stimulus property to which its response is specific (Fodor & Pylyshyn 1981). Accordingly, it is traditionally considered a further condition of cognitive penetration that the effect must sustain some semantic coherence or logical relation between the contents of perception and the contents of cognition (Macpherson 2012; Pylyshyn 1999).

Contrary to its prima facie plausibility, then, it is in fact hard to find a convincing empirical example of cognitive penetration, if by the term we mean:

> (CogPen) A phenomenally conscious experience is cognitively penetrated $=_{df}$
>
> (i) the effect is genuinely *perceptual*;
>
> (ii) the source of the effect is genuinely *cognitive*;
>
> (iii) the effect is *not mediated* by changes in the stimuli, the state of the sense organs, or the allocation of attention;
>
> (iv) the effect sustains some *semantic coherence* or *logical relation* between the contents of the penetrating cognitive state and the penetrated perceptual state.

It is against this backdrop that a particular set of psychological experiments has been singled out recently as providing possibly the most convincing case yet for cognitive penetration. In these experiments, subjects consistently judged images of prototypical White faces as lighter than equiluminant images of prototypical Black faces. Importantly, subjects also judged an image of a racially ambiguous face as lighter when it was labeled 'White' as opposed to 'Black' (Levin & Banaji 2006). It is especially on account of the latter result that these findings have been widely interpreted as demonstrating that even relatively abstract beliefs or concepts like those regarding race can influence our perceptual

experience of color or lightness (e.g., Hugenberg & Sacco 2008; Levin & Banaji 2006; Macpherson 2012; Stokes 2013; Vetter & Newen 2014).

If this claim is true, then it has enormous consequences for the debate on cognitive penetration, as well as for various issues in philosophy. So it is crucial to establish whether the experiments support such a claim. I argue that they do not. As a first step, let us then acquaint ourselves with the experiments in question.

## 2. THE EXPERIMENTAL EVIDENCE

In a set of four experiments Levin and Banaji (2006) tested whether images of faces categorized as White are perceived as lighter than equiluminant images of faces categorized as Black.

### *2.1. First Experiment: Judged Lightness of Prototypical White/Black Faces*

In the first experiment, subjects completed a series of trials in which they were presented a pair of computer-morphed grayscale images of prototypical White or Black faces. The faces could be either of the same race or different races. Within each trial, the initial luminance of the images was offset. The task was to adjust the luminance of one image to match the luminance of the other image.

As expected, subjects adjusted images to objectively lighter/darker levels when matching for White/Black faces. The effect was significant for all combinations of faces, but it was greater in the different-race trials than in the same-race trials.

### *2.2. Second Experiment: Judged Lightness of Racially Ambiguous Faces*

In the second experiment, subjects were divided into two groups. During instructions, one group was presented an image of a computer-morphed grayscale image of a racially ambiguous face labeled 'White,' next to an image of an equiluminant prototypical Black face labeled 'Black' ('White' ambiguous / unambiguous Black). The other group was presented the same ambiguous face but labeled 'Black,' next to an image of an equiluminant prototypical White face labeled 'White' ('Black' ambiguous / unambiguous White).





In the trials, subjects were presented one image at a time without labels, next to an adjustable gray patch. The initial luminance of the patch was offset. The task was to adjust the luminance of the patch to match the luminance of the presented face.

Consonant with previous findings, subjects adjusted the gray patch to an objectively lighter level for White as opposed to Black faces. Importantly, this was also the case for 'White' as opposed to 'Black' ambiguous faces. The magnitude of the effect was even greater than in the previous experiment.

### *2.3. Third Experiment: Judged Lightness of Line-Drawing Faces*

In the third experiment, the stimuli of the previous experiment were changed for evenly gray-filled dark and light line drawings of a Black, White, and ambiguous faces, respectively. Procedures were similar to those of the previous experiment.[2]

Cohering with previous findings, subjects adjusted a gray patch to an objectively lighter level for White as opposed to Black faces. This was also the case for ambiguous faces initially presented next to a Black as opposed to a White face. Though the magnitude of the effect was smaller than for computer-morphed images, importantly, both the magnitude and direction of the effect were similar for dark and light line drawings.

### *2.4. Fourth Experiment: Differentiation of Face Pairs by Race*

In the fourth experiment, akin to the first experiment, subjects were presented a pair of computer-morphed images of prototypical White or Black faces of either the same or different races. Subjects were told that the faces would vary in luminance, but they were instructed to ignore these differences. The task was to indicate by pressing a button as quickly and accurately as possible whether the presented faces were of the same or different races.

The results indicated that the more similar in luminance two faces were, the longer it took to discriminate the faces by race. Yet, importantly, reaction times were slowest not when two faces were objectively equiluminant, but when the Black face was a bit lighter than the White face.

### *2.5. Summary and Conclusion*

Hence, it would seem that differences in facial form (White vs. Black vs. ambiguous), lexical labeling ('White' vs. 'Black'), and/or visual context (ambiguous/White vs. ambiguous/Black) influence the judged lightness of faces. That at least some portion of this influence is genuinely perceptual is underscored by the finding that the discrimination of faces by race is involuntarily sensitive to differences in the luminance or lightness of those faces. Since the effect is apparently insensitive to line-drawing color (light vs. dark), it does not seem to be mediated by selectively attending to lighter/darker regions of White/Black faces. These considerations jointly suggest that, indeed, "the relative associations between lightness and White faces and between darkness and Black faces . . . make White and Black faces appear lighter and darker, respectively, than they actually are" (Levin & Banaji 2006, p. 501).

Of course, one may still question whether the experimental results reflect genuine perceptual effects. For example, recall that in the first experiment, subjects judged White/Black faces as lighter/darker even in the same-race trials. Yet in these trials, the reference and the matching faces differed *only* in their initial luminance. However distorted a subject may perceive the lightness of a reference face, one would expect that s/he perceives the lightness of a matching face with identical form features as similarly distorted. So the distortions should cancel out in the same-race trials, and subjects should adjust the luminance of the matching face to roughly the same objective level as that of the reference face. That this was not the case suggests that at least some portion of the effect may be better explained by perceptual misjudgments or response biases than genuine perceptual distortions (cf. the "El Greco fallacy"; Firestone & Scholl 2014).

Levin and Banaji attempt to explain away the above finding in purely perceptual terms. But we need not concern ourselves with their account. For even if their proposal is wrong, recall that the distortions were greater in the different-race trials than in the same-race trials (first experiment), and the effect was even greater when the matching stimulus was a mere gray patch (second experiment). Further, it seems very hard to explain on cognitive grounds why subjects would be slowest in discriminating a pair of faces by race not when the faces are equiluminant, but when the Black face is actually a bit lighter than the White face (fourth experiment). Given these considerations, the results of the same-race trials leave plenty of room for assuming that at least *some* portion of the effect was perceptual.



Some may still worry that the results reflect stimulus or attentional artifacts. For example, Machery (forthcoming) emphasizes that the eyes of the prototypical White face were objectively darker than the eyes of the prototypical Black face. This was needed to assure that the average luminance of the faces were identical. But recall that the effect was similar for evenly filled line-drawing faces, whether the lines were light or dark (third experiment). So even if some portion of the effect is contaminated by artifacts, perhaps another portion is not.

Note, though, that nothing ultimately hangs on the above points. For the crux of my argument is that *even if* the face lightness illusion is a genuine perceptual effect that is uncontaminated by artifacts, the effect *still doesn't* constitute a convincing case of cognitive penetration. If it turns out that the effect is not genuinely perceptual and/or it is mediated by artifacts, then so much the worse for a thesis of cognitive penetration.

## 3. AN UNWARRANTED ARGUMENT

What makes Levin and Banaji's findings so special? Note that it has long been known that color judgments are distorted for color-diagnostic objects. For example, in a classic psychological study conducted by Delk and Fillenbaum (1965), subjects judged orange-red cutout figures as more red for red-associated figures (heart, apple, lips) than for other-color-associated (horse, bell, mushroom) or non-color-associated figures (oval, circle, ellipse). Such results cohere with more recent evidence according to which the subjective gray point of color-diagnostic objects is generally shifted towards the complementary color of those objects. For example, subjects set the gray point of an image of a typically blue object like a Nivea tin to what is objectively somewhat yellow, and the gray point of an image of a typically yellow object like an UHU glue to what is objectively somewhat blue (Witzel et al. 2011).

Some recent studies question whether such findings reflect genuine perceptual phenomena (e.g., Gross et al. 2014, this volume). But assume that they do. Still, even defendants of cognitive penetration typ-

ically allow that such effects of perceptual learning may ultimately be explained by non-cognitive factors. For example, as Macpherson notes,

> It might be that the early visual system, autonomously from belief or other cognitive states, alters the colour experiences of characteristically red shapes, to make them appear more red than they really are. This might happen in accordance with associationist principles, so that it is past exposure to a certain shape having a certain colour that has altered the way the visual system processes that shape's colour. (2012, p. 46)

Now distortions in the judged lightness of prototypical Black/White faces seem to afford a similar explanation in terms of intra-visual shape-color (or form-lightness) associations. Since such intra-visual associations are plausibly cognitively impenetrable, this would explain why face discrimination by race is sensitive to differences in the luminance of faces *despite* explicit instructions to disregard such differences (and hence, why the face lightness illusion persists in the face of contrary beliefs and desires). Accordingly, in themselves, these findings of Levin and Banaji warrant no special attention.

Yet the finding that subjects judged the very *same* image of a racially ambiguous face as relatively lighter when it was labeled 'White' as opposed to 'Black' strike some as especially revealing of cognitive penetration. The apparent basis for this view is an unargued conviction that "categorization was clearly done at the cognitive level as it was the labelling of the face that was responsible for the effect" (Macpherson 2012, p. 48). Based on this reasoning, a rather straightforward argument for cognitive penetration seems to run as follows:

Argument from lexical labeling

(1) Distortions in the perceived lightness of ambiguous faces were a function of *lexical labeling*.

(2) If the relevant distortions were a function of lexical labeling, then the effect was mediated by *cognitive categorization*.

(3) If the effect was mediated by cognitive categorization, then low-level visual experience was *cognitively penetrated*.





*Hence*,

(4) Low-level visual experience was cognitively penetrated.

I claim that none of the assumptions of this argument are warranted, and hence there is little reason to assume that low-level visual experience is cognitively penetrable. Here's why.

### *3.1.* Re *(2): Lexical vs. Conceptual vs. Cognitive Effects*

Let us first consider premise (2), according to which a labeling effect entails a cognitive effect. Why assume this? The premise seems to rely on the following (hidden) assumptions:

(2a) If the effect was lexically mediated, then it was *conceptually* mediated.

(2b) If the effect was conceptually mediated, then it was *cognitively* mediated.

Yet these assumptions are hardly a matter of course. On the contrary, there may be various reasons to question both. In the following two subsections, I provide some considerations accordingly.

#### 3.1.1. *Re* (2a): Lexical-Visual Associations

In a classical experiment by Swinney (1979), subjects heard sentences that contained a semantically ambiguous word, such as:

> Rumor had it that, for years, the government building had been plagued with problems. The man was not surprised when he found several spiders, roaches, and other *bugs* in the corner of his room.

Simultaneously with the ambiguous word 'bug' — which can mean either insect or surveillance device — subjects were visually presented with a letter string. Their task was to decide as quickly as possible whether the letter string was a word or a non-word. As expected, lexical decisions were facilitated (only) for related words. Yet, importantly, this facilitation occurred *irrespective* of whether the related word was contextually relevant ('ant') or contextually inappropriate ('spy').

Such findings suggest that, at least in some cases, the effects of sentence context on lexical access do not have so much to do with semantic processing *per se*. Rather, the effects may be mediated by certain intra-modular associations between mere lexical forms stored in a proprietary database of a language module (Fodor 1983)). This coheres with Collins and Loftus's (1975) spreading-activation theory of semantic processing, according to which the names of concepts (words) are stored in a lexical network (dictionary) that is separate from — albeit connected with — a semantic network of concepts.

Now the hypothesis of intra-modular/inter-lexical associations is easily extended to inter-modular/cross-modal associations. Accordingly, it is at least an empirical possibility that the lightness distortion effect for ambiguous faces was mediated by certain *infra*-cognitive associations between lexical items like 'black' and 'white' stored in a language module, and some corresponding color representations (of, e.g., black and white) stored in a visual module. Associations of such a kind might imply some degree of cross-penetrability between modular systems. But, importantly, these systems may still be encapsulated from cognition at large. This coheres with evidence that cross-modal effects like the "McGurk effect" — whereby visual exposure to the lip movements of a speaker influences the perception of concurrent speech sounds (McGurk & MacDonald 1976) — are themselves undergirded by cognitively impenetrable processes (Fodor 1983, 1988; Pylyshyn 2003).

Note that I am not arguing that the above proposal in particular is correct. Indeed, I will later argue that the distortion effect for ambiguous faces was most likely not lexically mediated. Neither is my point that the relevant perceptual distortions were not conceptually mediated. My issue is specifically with the conditional according to which if the ambiguous-face lightness illusion was lexically mediated, then it *follows* that the effect was conceptually mediated.

Hence, my reason for referencing the above proposals is mainly to underscore that, contra (2a), there are various empirical and theoretical considerations on the basis of which one may wish to be cautious in inferring a conceptual effect from a lexical effect. This is a dialectically important point, for the burden of my present argumentation is to show that none of the (hidden) premises of the argument from labeling are as firmly grounded as proponents of the argument would have us believe.





Examples like those of Swinney, Fodor, Collins and Loftus, or McGurk and MacDonald provide at least some reasons for resisting (2a). But not much ultimately hangs on this, as I will presently argue that (2b) is itself unjustified.

#### 3.1.2. *Re* (2b): Non-Cognitive Concepts

According to (2b), if the relevant effect was conceptually mediated, then it follows that it was cognitively mediated. If this claim is analytic, then so be it. But it strikes me as far from clear that the claim is justified on a substantive reading.

For example, consider the original dialectic of the cognitive-penetration debate. The cognitive revolution in psychology led to a widespread belief that, as opposed to earlier theorizing, even relatively low-level perceptual processes are by and large inferentially mediated (within the usual limitations of the metaphor). Inferences need premises, and at the time, it seemed natural to many that these premises must be represented in the mind by sentence-like representations the constituents of which are concepts (Fodor 1975, 2008).

Traditional proponents of cognitive impenetrability accepted this view. What they vehemently denied is that it *follows* from the inferential nature of a mental process that it is also informationally promiscuous and hence cognitively penetrable (Fodor 1985; Pylyshyn 1984). Accordingly, proponents of impenetrability may allow that perceptual processes are by and large conceptually mediated, yet still deny that these processes are by and large cognitively influenced.

With that said, it is important to keep in mind that the issue of whether conceptual mediation entails cognitive mediation is orthogonal to the issue of whether perception is cognitively penetrable. For example, albeit a staunch defender of cognitive impenetrability, Raftopoulos (2012) agrees that there is a mutual entailment relation between conceptpood and cognitive penetrability. On the other hand, whereas Toribio (2014) seems neutral about cognitive penetrability, she argues that the mutual entailment thesis is true only if it is trivial, or if it completely fails to engage with the contemporary literature on perceptual nonconceptualism. I will certainly not resolve this issue here. My point is merely that, given such considerations and controversies, it is hardly a matter of course that one may infer a cognitive effect on the basis of

a conceptual effect.

Of course, a defender of cognitive penetration may still hold that such an inference is warranted in the particular case of the ambiguous-face illusion. For example, s/he might note that even if we allow for the existence of non-cognitive concepts, surely, concepts of race are still cognitive. Perhaps so; but nothing follows from this. For notice that the labels used by Levin and Banaji ('Black' and 'White') are ambiguous between color terms and race terms. So *even if* distortions in the perceived lightness of ambiguous faces were a function of lexical labeling, it would still be an open empirical possibility that the effect was mediated by non-cognitive concepts.

Indeed, "psychophysical" concepts of color are prime candidates of non-cognitive concepts that are available for tokening by cognitively impenetrable perceptual systems (Fodor 1987). So a skeptic of cognitive penetration may well grant that the relevant effect was mediated by low-level color concepts like BLACK or WHITE. Contra (2b), it is unclear why it would follow that the effects of these color concepts were in turn mediated by high-level cognitive race concepts like AFRICAN-AMERICAN or CAUCASIAN, respectively.

An anonymous referee noted that even if the above scenario is an empirically real possibility, it is still abductively more plausible, all things considered, that the ambiguous face lightness illusion was cognitively mediated. I disagree — but more on that later. The relevant reply is that I am here not concerned with whether the consequent of (2b) — the claim that there was a cognitive effect — is abductively plausible. The target of this subsection is the very conditional according to which a conceptual effect entails a cognitive effect. It is this claim, I have argued, that there is plenty of room to resist.

### *3.2.* Re *(3): Cognitive Priming vs. Cognitive Penetration*

So far, I have provided some reasons to resist taking it for granted that the assumed labeling effect was conceptually (2a) and hence cognitively mediated (2b). Accordingly, it would seem that the second premise (2) of the argument from labeling is at least more questionable than defenders of cognitive penetration suggest. Of course, this still leaves open the possibility that the ambiguous-face lightness illusion was, as a matter of empirical fact, cognitively mediated. According





to premise (3), it would then follow that low-level visual experience was cognitively penetrated. Is there reason to question this claim? I argue that indeed there is.

#### 3.2.1. Semantic/Logical Coherence

It is significant that the face lightness illusion for prototypical faces persists even if one knows that the faces are objectively equiluminant. Both Levin and Banaji, and Macpherson emphasize this. This is somewhat ironical given that the persistence of an illusion in the face of contrary beliefs is traditionally considered strong prima facie evidence that the very states or processes underserving the illusion are cognitively impenetrable.

Beliefs and similar attitudes are paradigmatically cognitively penetrable. So the persistence of the face lightness illusion strongly suggests that whatever states or processes underlie the face lightness illusion, beliefs and similar attitudes are not among them. Macpherson suggests accordingly that perhaps the illusion is not mediated by beliefs, but by merely primed cognitive concepts. But this suggestion cannot satisfy the condition of sustaining some semantic or logical coherence between the contents of cognition and the contents of perception. For priming is a reflex-like, non-inferential process, that is notoriously insensitive to the compositional semantics of syntactically structured representations in general, and logical operators in particular. Let me explain.

For example, consider an experiment in which subjects first watched as an experimenter poured sugar from a single source into two separate clean bottles. After this, the experimenter asked the subjects to place a label reading "not sodium cyanide, not poison" on a bottle of their choice, and a label reading "sucrose, table sugar" on the other bottle. Amazingly, subjects still rated drinks sweetened with sugar from the bottle labeled "not sodium cyanide, not poison" as less desirable, and they were also more reluctant to take a sip from these drinks (Rozin et al. 1990). This coheres with evidence on "negative suggestions" in hypnosis and advertising, whereby a person to whom it is suggested not to think or feel or do something, is as a result *more* likely to think or feel or do that thing, respectively (cf., e.g., Cyna & Lang 2010).

So, then, say a subject is shown an ambiguous face labeled 'White,' whereas another subject is shown the same face labeled 'not White.'

Will the two subjects perceive the face differently? The priming account predicts that the answer is no. The labels are expected to exert an identical effect (if any) on perception, in virtue of an identical priming of some concept corresponding to WHITE.

Now even on Macpherson's account, "perceptual experience is cognitively impenetrable if it is not possible for two subjects . . . to have two different experiences on account of a difference in their cognitive systems" (2012, p. 29). In our example, one subject believes that the ambiguous face is White, and hence only tokens WHITE. The other subject believes that the face is non-White, and hence also tokens NOT WHITE. These are logically contradictory beliefs, with correspondingly different concepts employed. If the priming account predicts that such differences have no bearing on perceptual experience, then the account would not seem to satisfy the condition of semantic/logical coherence. Hence, contra (3), the mere influence of some cognitive category does not entail cognitive penetration.

#### 3.2.2. "Weak" Cognitive Penetration

Granted, it would still be extremely interesting if one's beliefs could indirectly influence perceptual experience via the automatic priming of cognitive concepts. Accordingly, a defender of cognitive penetration may perhaps simply drop or at least modify the semantic/logical condition in such a way as to allow for cognitive priming effects. This would result in a weaker thesis of cognitive penetration than is of interest here. All the same, I don't think there are good reasons to assume that low-level visual experience is even "weakly" cognitively penetrable.

For example, say I believe that a grayscale image of a face I am presented depicts a Chinese person. As it happens, I associate China with Communism, and Communism with the color red. Thus, when my concept CHINESE is primed, RED is correspondingly primed. Will I then perceive the face as somewhat red?

That the very question seems absurd suggests that whatever processes underlie the face lightness distortion effect, those processes are stimulus-triggered and modality-specific. In turn, this suggests that the effect is not mediated by relatively abstract concepts like those of race. I will soon argue accordingly that the effect is more plausibly mediated by relatively low-level perceptual representations of shape (facial form) and color (lightness).



Yet one may perhaps object that my example is misleading because, as opposed to color associates of race, color associates of political/socioeconomic systems or ideologies are irrelevant to perceiving the lightness of a face. Indeed they are. But how could the processes that mediate priming take this into account? In particular, how could reflex-like, non-inferential processes that are insensitive to structured/composed representations and rationality constraints like semantic/logical coherence assess for and selectively prime concepts based on relevance?

Alternatively, a proponent may object that no one actually holds that it is primed concepts *per se* that penetrate perceptual experience. Perhaps the assumption is that priming can influence the level or threshold of activation of concepts that uncontroversially contribute to experience *when* activated. For example, perhaps the priming of a race concept like AFRICAN-AMERICAN influences visual experience by enhancing the activation level of a color concept like BLACK. But, the objection might go, whereas it is independently plausible that BLACK is activated by a grayscale image of a racially ambiguous face, there is little corresponding reason to assume that a *grayscale* image of a Chinese face leads to a stimulus-dependent activation of RED. So even if RED is primed by COMMUNISM upon exposure to a Chinese face, insofar as RED is only primed and not eventually activated, proponents of the priming account need not assume that the face is perceived as somewhat red.

But say the Chinese face is depicted not in grayscale, but in orange-red color, like the cutout figures of Delk and Fillenbaum. If it is plausible that grayscale images can elicit stimulus-dependent activation of color concepts like BLACK or WHITE, then it seems correspondingly plausible that orange-red images may elicit stimulus-dependent activation of color concepts like ORANGE or RED. And if it is plausible that the priming of AFRICAN-AMERICAN can influence visual experience via a modification of the actual activation level of BLACK, then it should seem correspondingly plausible that the priming of COMMUNISM may influence visual experience via a modification of the activation level of RED.

I know of no empirical study to date that rules out the above possibility *per se*. But I doubt anyone would hold their breath. If perceptual experience were really so thoroughly penetrable as the priming account

suggests, then it would be simply a mystery how, after nearly seven decades of research starting with Bruner & Goodman (1947), defenders of cognitive penetration still haven't provided a more convincing empirical demonstration than the Levin and Banaji effect (cf. Firestone & Scholl forthcoming; Machery forthcoming).

Hence, even if not conclusive, considerations of the above kind at least raise serious doubts about the empirical plausibility of "weak" cognitive penetration. In turn, this provides further reason to question the antecedent of premise (3), according to which the face lightness illusion was mediated by cognitive categories. In the previous subsection, I already noted that the illusion is unlikely to be undergirded by belief-like cognitive states. As we have now seen, the idea of conceptual penetration via cognitive priming is no more plausible. So my worry is not merely that we may lack good reason to assume that the relevant effect was cognitive (∼assume C). If my analysis is correct, there is also considerable reason to assume outright that this was in fact not the case (assume ∼C).

Note that, as opposed to the truth-functional account of material implication, there is broad consensus that it doesn't follow from a false antecedent that an indicative conditional regarding some empirical matter is true. So reasons to doubt the antecedent of (3) hardly translate to reasons to assume that the conditional as such (cognitive categorization ⊃ cognitive penetration) is true. Indeed, I already argued that the conditional is at best false on a standard understanding of cognitive penetration. On a weaker understanding, the claim may be either true or false, or simply lacking in truth value. Given these contingencies and the relative implausibility of the priming account, (3) could hardly be further from trivial.

### *3.3.* Re *(1): Lexical Labeling vs. Visual Context*

The above considerations notwithstanding, some may argue that a cognitive account of the ambiguous-face illusion is still abductively more plausible if the effect was indeed mediated by lexical labeling. So was it? Defenders of cognitive penetration seem to take it for granted that it was.[3] Premise (1) states as much. Yet a careful consideration of Levin and Banaji's experimental design provides plenty of space for doubt.

Recall that in the crucial (second) experiment, faces were only labeled during the instruction phase, at which point an ambiguous face always appeared next to an unambiguous Black or White face. After instructions, all faces were presented without labels and on their own. So whenever a face was labeled, it appeared in the visual context of another face; and whenever a face was unlabeled, it appeared on its own. In technical terms, this means that the factor of lexical labeling was confounded with the factor of visual context. This is quite a methodological flaw if the authors' intention was to test for the effects of lexical labeling.

Now Levin and Banaji advertise in their abstract that judgments of lightness were distorted "even for racially ambiguous faces that were disambiguated by labels" (501). This suggests (albeit it doesn't imply) that the crucial effect was indeed lexically mediated. Yet the authors are themselves more cautious when they claim in a later section of their paper that the ambiguous faces were differentiated "on the basis of their context *and*/*or* a label" (510; emphasis added). Indeed, one of the authors has noted in personal e-mail correspondence that s/he "would be surprised if the absence of a label had much of an effect." These remarks suggest that the effect may have been mediated (exclusively) by visual context.

Hence, for all we know, the relevant perceptual effect might be preserved if ambiguous faces are presented without labels in the context of unambiguous faces (no labels / visual context). On the other hand, it is a real empirical possibility that the effect would disappear if labeled ambiguous faces were presented on their own (labels / no visual context). These possibilities suggest not only that the effect may not have been mediated by labeling *alone*. More importantly, they suggest that the labels may have had *nothing* to do with the effect.

So, it would seem that not even the very starting premise (1) of the argument from labeling enjoys the support that one would have expected. That the mentioned design flaw has so far gone unnoticed is especially surprising given that the labeling effect is widely considered to be *the* prima facie most convincing piece of evidence in favor of cognitive penetration. Considering this oversight, it is really unclear whether the argument from labeling has enough wind left to convince any skeptics.

## 4. A TENTATIVE ACCOUNT OF THE PHENOMENA

So far, I have provided various reasons to doubt that the ambiguous-face lightness illusion is an example of cognitive penetration. In the last section, I noted in particular that given the design of the relevant experiment, it is a real empirical possibility that whereas a lexical label is neither necessary nor sufficient, a visual context of an unambiguous face is both necessary and sufficient for the effect to occur for ambiguous faces. How could this be so? In this section, I provide a sketch of a positive account.

### *4.1. Form-Lightness Associations*

Firstly, consider that the primary visual diagnostic feature of race is facial form. For example, Black faces tend to have thicker lips and a wider nose than White faces, and the features of racially mixed or ambiguous faces are typically somewhere in between. Correspondingly, Black faces tend to have darker skin tone than White faces, and the skin tone of racially mixed or ambiguous faces is typically somewhere in between. So why assume that the face lightness illusion is mediated by extra-perceptual categories of race? A much simpler and more plausible assumption is that the effect is mediated by intra-perceptual associations between facial form and color/lightness.

Despite their earlier remarks and suggestions to the contrary, Levin and Banaji turn out to argue exactly along these lines:

> This distortion might be considered a case where a set of correlated features mutually facilitate each other such that the presence of most members of the set causes activation of representations of the missing members. So, the correlation between *form* and shading causes shading features to be activated in the presence of *form* features.... Thus, Black faces might appear relatively dark...as the result of feature activations resulting from a *perceptual* classification. (511; emphasis added)

This assumption coheres well with evidence on the "memory color effect," of which the face lightness illusion is ultimately just a special case. The effect involves a modulation by memory colors of the perceived color of subjectively color-diagnostic objects. For example, recall that subjects seem to perceive a heart shape as more red (Delk & Fillenbaum 1965), or a Nivea tin as more blue (Witzel et al. 2011), than similarly colored objects of which these colors are not diagnostic. Though these effects seem to be object-sensitive, they are also apparently experience-driven, modality-specific, and thought-insensitive. The latter features militate against the idea that the memory color effect is cognitively mediated (Deroy 2013).

The Witzel et al. study is instructive in this sense. The authors only used images of objects that met pre-set criteria of high subjective color diagnosticity, as measured by reaction time and accuracy of identifying the typical color of an object. Yet out of the 14 images used, only 10 elicited a relevant observable effect; for only 7 images was the effect statistically significant; and even among these cases, there was great variation in the magnitude of the effect.

Even more importantly, the direction of the effect was effectively reversed for some images. In general, the subjective gray point of an object shifted toward the *complementary* color of the color associated with that object; e.g., toward yellow for a typically blue object like a Nivea tin. One would thus expect that the subjective gray point of typically red objects shifts toward green. Yet this was not the case. For red-associated objects like a heart shape or a Coke bottle, the subjective gray point shifted toward the *associated* color, i.e., red.

It is very hard to think of a plausible explanation as to why cognition would penetrate perception in such diverse and unexpected ways. On the other hand, a further finding that shifts in subjective gray point were particularly large for objects the typical colors of which are close to the daylight axis — an axis that passes from the bluish to the yellowish part of the color space — only underscores that, rather than the influence of cognition, these effects reflect built-in constraints of our perceptual system.

So there is good reason to assume that the associations between shape and color/lightness on which the memory color effect depends are intra-perceptually mediated. That the Levin and Banaji effect in particular is undergirded by intra-perceptual associations is further supported by a fresh study in which images of the prototypical White and Black faces were blurred to the extent that most subjects could not tell

the race of the faces. All the same, these subjects also judged the White face as relatively lighter than the Black face (Firestone & Scholl forthcoming). Though this finding does not rule out that perhaps race concepts were still unconsciously tokened or primed, it does provide further reason to question whether the cognitive-penetration account is all that plausible.

### *4.2. Lightness Contrast*

The hypothesis of intra-perceptual form-lightness associations thus predicts a memory color effect whereby White faces are perceived as lighter than equiluminant Black faces, and the lightness of ambiguous faces is perceived as being somewhere in between. If this hypothesis is correct, then presenting an ambiguous face next to a Black or White face is like presenting a shade of gray next to a shade of black or white, respectively. Yet it is a classic textbook example of lightness contrast that the very same shade of gray looks lighter/darker in the context of a black/white shade! So given a genuine memory color effect for faces, it should not be that surprising if the lightness of an ambiguous face is perceived differently in the context of a Black and a White face.

Note that perceptual contrast effects are a function of the perceived (subjective) and not the actual (objective) reflective properties of the stimuli. So it is a non-issue that the ambiguous and unambiguous faces were objectively identical in luminance. Nor is it an issue that, after instructions, in the trials in which subjects actually judged for lightness, the faces were presented one at a time. For anchoring effects due to simultaneous contrast in the instruction phase might have easily carried over to the first trial. After that, perceptual anchoring and adaptation might both explain successive (between-trial) contrast effects.

Hence, I propose that differences in the judged lightness of ambiguous faces had not so much to do with lexical labeling or racial disambiguation, as with lightness contrast mediated by intra-perceptual form-lightness associations. This hypothesis makes sense of various findings that would be very hard to explain on cognitive terms. For example, it tends to go without mention that in the labeling experiment, lightness judgments were only distorted for the apparently lighter face among an ambiguous/unambiguous face pair, and hence the distortions only tended toward over-lightening. So if an ambiguous 'White' face





was paired with an unambiguous Black face, the ambiguous face was judged as relatively lighter, whereas the unambiguous face was correctly judged. Yet if an ambiguous 'Black' face was paired with an unambiguous White face, the ambiguous face was correctly judged, and it was the unambiguous face that was judged as relatively lighter.

My proposal can easily account for this apparent anomaly. For example, lightness contrast may have led to an application by the perceptual system of a highest-luminance rule (Adelson 2000; Gilchrist et al. 1999), whereby the apparently lighter face was anchored to white. Of course, not much hangs on whether this is the actual rule or mechanism underlying the effect. The important point is that insofar as the relevant distortions are biased in such unexpected albeit consistent ways, an intra-perceptual account of the phenomena is indeed much simpler and more plausible than a cognitive account.

## 5. SUMMARY AND CONCLUSION

The conclusion of this paper is that the Levin and Banaji effect is an unconvincing case of cognitive penetration. While a prima facie powerful argument suggests the contrary based on the assumption that differential lightness judgments for ambiguous faces were a function of lexical labeling, on closer scrutiny, it turns out that the premises of this argument are neither empirically nor theoretically very well supported. Accordingly, I argued that the phenomena are more easily and plausibly explained by intra-perceptual than by cognitive mechanisms.

My account is clearly more parsimonious than a cognitive account, for it does not assume the involvement of any lexical or cognitive representations. It is also empirically more plausible insofar as it does not posit any as-of-yet unestablished or controversial psychological mechanisms. And it is abductively preferable insofar as it can explain everything that a cognitive account can, and also more. For example, my proposal can easily account for why the face lightness illusion persists in the face of contrary beliefs, or why the effect was selective and unidirectional in the ambiguous-face experiment.

The upshot is that memory color effects like the face lightness distortion effect might be sufficiently explained by the internal workings of cognitively impenetrable perceptual systems. This suggests that

perceptual systems are more plastic and hence psychologically richer than is commonly assumed. Importantly, and contrary to a currently widespread view in philosophy, it also suggests that the significance of cognitively impenetrable systems is not confined to a subpersonal or unconscious level.

For example, consider the epistemic threat posed by cognitive penetration with respect to the role of perception in justification. If our beliefs can penetrate our perceptual experience, which we in turn consider as evidence for our beliefs, then the structure of belief formation would seem to be circular (Siegel 2011). If my proposal is correct, and perceptual experience is cognitively impenetrable, then this worry might be overcome.

It bears emphasis that not all accounts of cognitive impenetrability circumvent the threat of circularity. For example, a common strategy of explaining away alleged effects of cognitive penetration is in terms of perceptual misjudgment. One might thus maintain that perceptual experience remains intact in the face of cognitive biases in judgment. But if a subject is genuinely deluded about his or her perceptual experience, then, akin to the case of cognitive penetration, she might just as well consider her mistaken judgments as evidence for beliefs that were the source of the bias in the first place.

An apparent further advantage of my account, then, is that it can circumvent the epistemic threat of circularity without the undesirable implication that we are possibly way more often deluded about our first-person phenomenal experiences than epistemologists and psychologists would have it (not to mention regular folk). Of course, if my proposal is correct, we might still often be victims of perceptual illusions. So certain epistemic worries remain. But at least these are not the worries of circular belief formation or introspective bias.

Now one may wonder whether or how my analysis extends beyond, for example, those of Deroy (2013) or Firestone & Scholl (forthcoming). Deroy provides a thorough analysis of the memory color effect, but she doesn't address the Levin and Banaji effect. Firestone and Scholl address the effect head on, but only for prototypical White/Black faces.

As I noted earlier, most defenders of cognitive penetration are in principle prepared to grant that standard memory color effects (e.g. that hearts are perceived as more red) are intra-perceptually mediated. These defenders may be correspondingly ready to grant that the face lightness illusion for prototypical White/Black faces affords a non-cognitive explanation. Yet these defenders still argue that the illusion for ambiguous faces cannot be so explained. To the best of my knowledge, I am the first to expose the weaknesses of this argumentation, as I am also the first to offer a plausible low-level account of the crucial evidence.

Of course, debunking evidence or arguments that defenders don't find crucial is still important progress. But it is unlikely to make serious headway in the overall debate. So an evident dialectic advantage of my analysis is that it specifically addresses what many defenders themselves pinpoint as the most convincing case for cognitive penetration. Given this dialectic, my challenge cannot be swept aside by noting that I have merely cast doubt on a single case. Hence, if my analysis of the relevant evidence and the argument that builds on it is more or less correct, the burden is indeed on defenders to argue why anyone should still assume that low-level visual experience is cognitively penetrable.



## Notes

[1] For more liberal views on attention-mediated effects, see Macpherson (2012) or Prinz (2006).

[2] As opposed to the between-subjects design of the second experiment, though, the third experiment had a within-subjects design. So all subjects completed trials both for when the ambiguous face initially appeared next to an unambiguous White face, and for when the ambiguous face initially appeared next to an unambiguous Black face. Unfortunately, Levin and Banaji do not mention whether they used labels in this experiment.

[3] For example, Macpherson describes the crucial (second) experiment as follows: "a racially ambiguous face was labelled either as the face of a white person or the face of a black person and this factor alone determined what shade of grey the subjects chose as a match for the lightness of the face" (2012, p. 48).

## References

Adelson, E. H. 2000. 'Lightness Perception and Lightness Illusions'. In M. Gazzaniga (ed.) 'The New Cognitive Neurosciences', 339–351. Cambridge, MA: The MIT Press, 2nd ed.

Bruner, J. S. & Goodman, C. C. 1947. 'Value and need as organizing factors in perception.' *Journal of Abnormal and Social Psychology* 42: 33.

Collins, A. M. & Loftus, E. F. 1975. 'A spreading-activation theory of semantic processing'. *Psychological Review* 82, no. 6: 407.

Collins, J. A. & Olson, I. R. 2014. 'Knowledge is power: How conceptual knowledge transforms visual cognition.' *Psychonomic Bulletin & Review* 21, no. 4: 843–860.
Cyna, A. M. & Lang, E. V. 2010. 'How words hurt'. In A. M. Cyna, M. I. Andrew, S. G. M. Tan & A. F. Smith (eds.) 'Handbook of Communication in Anaesthesia & Critical Care: A Practical Guide to Exploring the Art', 30–37. New York: Oxford University Press.
Delk, J. L. & Fillenbaum, S. 1965. 'Differences in perceived color as a function of characteristic color'. *The American Journal of Psychology* 78, no. 2: 290–293.
Deroy, O. 2013. 'Object-sensitivity versus cognitive penetrability of perception'. *Philosophical Studies* 162: 1–21.
Firestone, C. & Scholl, B. forthcoming. 'Can you experience 'top-down' effects on perception?: The case of race categories and perceived lightness'. *Psychonomic Bulletin & Review*.
Firestone, C. & Scholl, B. J. 2014. '"Top-down" effects where none should be found: The El Greco fallacy in perception research'. *Psychological Science* 25, no. 1: 38–46.
Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.
——. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: The MIT Press/Bradford Book.
——. 1985. 'Précis of 'The modularity of mind''. *Behavioral and Brain Sciences* 8: 1–5.
——. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: The MIT Press/Bradford Book.
——. 1988. 'A Reply to Churchland's" Perceptual Plasticity and Theoretical Neutrality'". *Philosophy of Science* 55, no. 2: 188–198.
——. 2008. *LOT 2: The Language of Thought Revisited: The Language of Thought Revisited*. Oxford University Press.
Fodor, J. A. & Pylyshyn, Z. W. 1981. 'How direct is visual perception?: Some reflections on Gibson's "ecological approach"'. *Cognition* 9, no. 2: 139–196.
Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., Spehar, B., Annan, V. & Economou, E. 1999. 'An anchoring theory of lightness perception'. *Psychological Review* 106, no. 4: 795–834.
Gross, S., Chaisilprungraun, T., Kaplan, E., Menendez, J. & Flombaum, J. 2014. 'Problems for the purported cognitive penetration of perceptual color experience and Macpherson's proposed mechanism'. This volume.
Hugenberg, K. & Sacco, D. F. 2008. 'Social categorization and stereotyping: How social categorization biases person perception and face memory'. *Social and Personality Psychology Compass* 2, no. 2: 1052–1072.
Levin, D. T. & Banaji, M. R. 2006. 'Distortions in the perceived lightness of faces: The role of race categories'. *Journal of Experimental Psychology: General* 135, no. 4: 501–512.
Machery, E. forthcoming. 'Cognitive penetrability: A no-progress report'. In J. Zeimbekis & A. Raftopoulos (eds.) 'Cognitive Penetrability'. Oxford University Press.
Macpherson, F. 2012. 'Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism'. *Philosophy and Phenomenological Research* 84, no. 1: 24–62.
McGurk, H. & MacDonald, J. 1976. 'Hearing lips and seeing voices'. *Nature* 264, no. 5588: 746–748.
Prinz, J. 2006. 'Is the mind really modular?' In R. J. Stainton (ed.) 'Contemporary Debates in Cognitive Science', 22–36. Malden: Blackwell Publishing.
Pylyshyn, Z. 1999. 'Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception'. *Behavioral and Brain Sciences* 22, no. 03: 341–365.
Pylyshyn, Z. W. 1984. *Computation and Cognition: Towards a foundation for cognitive science*. Cambridge, MA: The MIT Press/Bradford Book.
——. 2003. *Seeing and Visualizing: It's Not What You Think*. Cambridge, MA: The MIT Press/Bradford Book.
Raftopoulos, A. 2001. 'Is perception informationally encapsulated?: The issue of the theory-ladenness of perception'. *Cognitive Science* 25, no. 3: 423–451.
——. 2012. 'The cognitive impenetrability of the content of early vision is a necessary and sufficient condition for purely nonconceptual content'. *Philosophical Psychology* 5, no. 5: 601–620.
Ramachandran, V. S., Rogers-Ramachandran, D. & Cobb, S. 1995. 'Touching the phantom limb'. *Nature* 377, no. 6549: 489–490.
Rozin, P., Markwith, M. & Ross, B. 1990. 'The sympathetic magical law of similarity, nominal realism and neglect of negatives in response to negative labels'. *Psychological Science* 1, no. 6: 383–384.
Siegel, S. 2011. 'Cognitive penetrability and perceptual justification'. *Noûs* 46, no. 2: 201–222.
Stokes, D. 2013. 'Cognitive penetrability of perception'. *Philosophy Compass* 8, no. 7: 646–663.
Swinney, D. A. 1979. 'Lexical access during sentence comprehension: (Re)consideration of context effects'. *Journal of Verbal Learning and Verbal Behavior* 18, no. 6: 645–659.
Toribio, J. 2014. 'Nonconceptualism and the cognitive impenetrability of early vision'. *Philosophical Psychology* 27, no. 5: 621–642.
Vetter, P. & Newen, A. 2014. 'Varieties of cognitive penetration in visual perception'. *Consciousness and Cognition* 27: 62–75.
Witzel, C., Valkova, H., Hansen, T. & Gegenfurtner, K. R. 2011. 'Object knowledge modulates colour appearance'. *i-Perception* 2, no. 1: 13.